Stress-testing Fairness Mitigation Techniques under Distribution Shift using Synthetic Data

Karan Bhanot
bhanotkaran22@gmail.com
Rensselaer Polytechnic Institute
Troy, New York, USA

Ioana Baldini
ioana@us.ibm.com
IBM Research
Yorktown Heights, New York, USA

Dennis Wei
dwei@us.ibm.com
IBM Research
Yorktown Heights, New York, USA

Jiaming Zeng
jiaming@ibm.com
IBM Research
Cambridge, Massachusetts, USA

Kristin P. Bennett
bennek@rpi.edu
Rensselaer Polytechnic Institute
Troy, New York, USA

ABSTRACT

Machine learning (ML) models may suffer from biases against certain subgroups defined by protected attributes. ML fairness mitigation techniques aim to resolve these biases. But how well do these fairness methods work in practice? To date, their evaluation has largely been limited to a few overly-used datasets that do not consider distribution shifts. In this paper, we propose the design of an "auditor" that uses synthetic data generation to create a grid of scenarios with distribution shifts to stress-test these techniques. We provide an explanation of the design of this auditor along with fairness audits of the reweighing method for fairness mitigation, using synthetic versions of MIMIC-III and Adult Income datasets. The paper also highlights the importance and potential benefits of doing fairness auditing of algorithms.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Machine learning**.

KEYWORDS

machine learning fairness, stress-testing, distribution shift, fairness mitigation, synthetic data

ACM Reference Format:
Karan Bhanot, Ioana Baldini, Dennis Wei, Jiaming Zeng, and Kristin P. Bennett. 2018. Stress-testing Fairness Mitigation Techniques under Distribution Shift using Synthetic Data. *ACM Trans. Graph.* 37, 4, Article 111 (August 2018), 7 pages. https://doi.org/XXXXXXX.XXXXXXX

Authors' addresses: Karan Bhanot, bhanotkaran22@gmail.com, Rensselaer Polytechnic Institute, Troy, New York, USA; Ioana Baldini, ioana@us.ibm.com, IBM Research, Yorktown Heights, New York, USA; Dennis Wei, dwei@us.ibm.com, IBM Research, Yorktown Heights, New York, USA; Jiaming Zeng, jiaming@ibm.com, IBM Research, Cambridge, Massachusetts, USA; Kristin P. Bennett, bennek@rpi.edu, Rensselaer Polytechnic Institute, Troy, New York, USA.

0730-0301/2018/8-ART111 $15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Data-driven approaches based on Machine Learning (ML) enable insights and deployment of solutions for real-world problems such as job hiring, screening candidates for college admission, granting loans and more [Makhlouf et al. 2021]. Such models learn patterns from historical data and use them to make decisions on unseen data, aiding professionals in making quicker and more informed decisions. However, recent research has revealed that such data and the models developed on them are often unfair [Buolamwini and Gebru 2018; Shankar et al. 2017], discriminating against certain groups over others. For example, the data in one of the most popular image datasets, ImageNet, disproportionately represents different countries, where 45% data is from the United States while only 3.1% data comes from China and India [Shankar et al. 2017]. In another study, the authors found that commercially available facial recognition systems often misclassify dark-skinned women more than light-skinned males [Buolamwini and Gebru 2018].

To rectify these biases, fairness mitigation techniques are designed to remove unwanted bias from the data and/or the models developed on them. Depending on their point of application in the ML pipeline, these techniques are categorized into: (a) preprocessing (e.g. reweighing [Calders et al. 2009] available in the toolkit of [Bellamy et al. 2018]), (b) in-processing (e.g. reductions [Agarwal et al. 2018]) and post-processing (e.g. reject option classification [Kamiran et al. 2012]). While many techniques exist, their applicability across a wide range of datasets and distribution shifts is often not investigated.

Recent literature reveals a growing interest in such an evaluation. One study proposed the learning of models with fairness considerations under different source and target distributions with missing protected attributes [Coston et al. 2019]. Another study highlighted the causal aspect of fairness properties under a distribution shift between training and deployment [Schrouff et al. 2022]. In contrast, we propose to evaluate fairness mitigation techniques across a range of distribution shifts using synthetic data, especially under limited dataset availability. Such a thorough evaluation is essential, as using these techniques across shifts without robust evaluation can lead to unpredictable performance and may even be ethically problematic, especially in high-impact domains such as healthcare. For example, if a fairness mitigation technique is chosen because it works well for Hospital A with certain demographics and is used without a fairness evaluation in Hospital B with different

demographics, then the technique might in fact fail to generalize as its applicability in the new distribution is unknown.

We propose the development of an "auditor" to stress-test fairness mitigation techniques across a variety of compound shifts, which combine multiple distribution shifts in attributes and outcomes. While datasets commonly used for fairness evaluation are already limited, the problem is worsened in healthcare, where access to data is often limited by privacy laws like Health Insurance Portability and Accountability Act (HIPAA) [Cntrs for Disease Cont and Prev 2018] and General Data Protection Regulation (GDPR) [European Parliament and of the Council (2016, Apr. 27) 2016]. Thus, we propose to augment limited existing datasets by generating synthetic data to represent a grid of scenarios with varying distribution shifts from the given dataset. Each such synthetically generated scenario corresponds to a shift where the fairness mitigation technique should be evaluated. This is a step towards the robust evaluation of current and future fairness mitigation techniques, ensuring their ethically responsible application across real-world scenarios.

In this paper, we show the strength of such an auditor by discussing preliminary results of the performance of a pre-processing technique, reweighing [Calders et al. 2009], on several synthetic scenarios generated using bootstrapping of two datasets: (a) MIMIC-III dataset [Johnson et al. 2016] and (b) Adult Income dataset [Kohavi and Becker 1996]. Extensions to other mitigation techniques, data generation methods and base datasets provide rich opportunities for future work. We discuss the observed results and highlight scenarios where the fairness mitigation technique struggled to remove bias with extended results in the appendix.

2 METHODOLOGY

2.1 Mathematical Formulation

Let A, X and Y represent three random variables such that A is one or more protected attributes (such as gender, race, ethnicity), X is the remaining set of attributes (such as blood pressure, heart rate) and Y is the outcome variable (such as presence or absence of a disease). For simplicity, we consider A to be a binary protected attribute such that $A = 1$ represents one group while the rest of the population is described by $A = 0$. Similarly, Y is also assumed to be a binary random variable. Thus, a dataset $D = \{A, X, Y\}$ can be described by the three variables representing a prediction task. Further, let $\hat{Y}$ represent the predicted outcome of a Machine Learning (ML) model generated on the dataset.

Table 1: Probability Distribution Table for Protected Attribute A and Outcome Y

		Protected Attribute		
		A = 0	A = 1	
Outcome	Y = 0	P(A = 0, Y = 0)	P(A = 1, Y = 0)	P(Y = 0)
	Y = 1	P(A = 0, Y = 1)	P(A = 1, Y = 1)	P(Y = 1)
		P(A = 0)	P(A = 1)	

Given our assumptions of binary A and Y, they each have a Bernoulli marginal distribution and a joint distribution given by the 2×2 table in Table 1, where the rows and columns sum to the corresponding marginal probabilities. This probability table is the basis for the distribution shifts considered in this paper. It is straightforward to extend our approach to multinomial A and Y.

2.2 Distribution Shift Scenarios

To generate various scenarios as a proxy for real-world shifts, we introduce distribution shifts to an existing dataset. There are four types of distribution shifts [Schrouff et al. 2022]: (a) Demographic shift, (b) Covariate shift, (c) Label shift and (d) Compound shift.

Amongst the four, compound shifts commonly occur in the real-world, for example in the healthcare setting of [Schrouff et al. 2022] due to different data sources for training versus deployment. Here we focus on compound shifts which consist of simultaneous demographic and label shifts in the data. We jointly change the proportions of the protected attribute A and outcome Y, while preserving $P(X = x|A = a, Y = y)$, the marginal density of X given A and Y.

To produce compound shifts, we make use of Iterative Proportional Fitting (IPF) which has been used across a variety of applications including estimation of resident characteristics, scaling population counts, estimating missing data and more [Lomax and Norman 2016]. IPF is a procedure to modify a two-dimensional probability table such that the row and column sums closely match new marginal probabilities [Norman 1999]. Based on Table 1, we define a starting contingency table for a given dataset D in terms of its corresponding counts (of records with $A = 0$, $Y = 0$, and so forth). We then take as input the new marginal row and column sums for A and Y respectively and use IPF to obtain the new cell values. This involves iteratively updating the cell values such that the new row and column sums closely match the marginals specified. These new cell values can then be used for generating specific datasets, resulting in compound shifts from the real data.

2.3 Auditor Design

Fairness mitigation techniques are often evaluated only on a small number of datasets. As a result, such evaluations are not robust and thus, the technique's applicability is not completely measured. This is troublesome as such techniques can be used in any domain or application where their effectiveness is unknown, while practitioners may believe that they are effective. Toward addressing this, we propose to use synthetic data generation along with IPF to create scenarios with distribution shifts from given datasets. For example, if the real data represents data from Hospital A, the synthetically generated scenario could represent a Hospital B with different demographics. This process shall then be repeated to create k different scenarios, each representing a different shift in demographic. Concretely, we propose an "auditor" as a pipeline for evaluating fairness mitigation techniques under such distribution shifts. The auditor generates a grid of scenarios with possible shifts and then performs a grid search across these scenarios to highlight the limitations of the technique. The proposed design for the auditor is shown in Figure 1. The auditor is designed to conduct a robust evaluation for any given technique with the steps described below:

- **Step 1:** Select dataset $D = \{A, X, Y\}$ to be used for evaluation. We create the starting contingency table for the protected attribute A and the outcome Y.
- **Step 2:** We define various combinations of the proportions of protected attribute A and outcome Y. These become the new

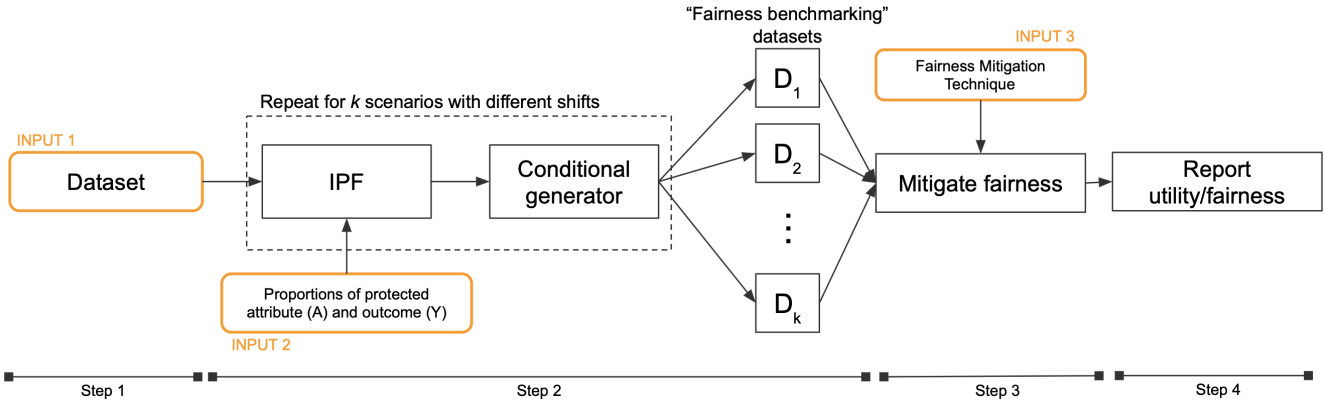

Figure 1: Proposed auditor design for robust evaluation of fairness mitigation techniques. The auditor takes three inputs (orange boxes): (a) Dataset, (b) Proportions of the protected attribute and outcome and (c) Fairness Mitigation Technique. The auditor synthetically generates a grid of k scenarios with distribution shifts and evaluates the mitigation technique against them. The utility-fairness results are reported.

marginal totals for the contingency table and constraints for IPF. After running IPF, a conditional synthetic data generator is then used to generate data conditioned on each cell (e.g. $A = 0$ and $Y = 0$) and with the number of samples found using IPF. This is repeated for different compound shifts, resulting in k different datasets $D_1, D_2, D_3 ... D_k$, each representing a synthetic scenario in a grid of possible scenarios. These are referred to as "fairness benchmarking" datasets.

- **Step 3:** The fairness mitigation technique is then applied to all these synthetic datasets in an attempt to reduce bias as measured by a user-specified fairness metric f. Based on the type of mitigation technique, the mitigation is performed before or after ML model training (e.g. Random Forest) for outcome prediction $\hat{Y}$.
- **Step 4:** The resulting fairness and utility scores are reported to the user highlighting scenarios where the fairness mitigation struggled to remove bias.

The auditor is proposed as a robust mechanism to stress-test fairness mitigation techniques under distribution shifts, to ensure the applicability of a technique under different scenarios. The evaluation can further be made more comprehensive by including additional datasets and protected attributes.

2.4 Experiment Description

We demonstrate the effectiveness of the proposed auditor by considering two case studies. The first case study explores a subset of the commonly used healthcare dataset, MIMIC-III [Johnson et al. 2016]. The source data is derived from MIMIC-III to replicate a previously published study [Bhanot et al. 2022]. The dataset includes several attributes with the outcome variable indicating whether someone dies within 30 days of hospital ICU admission or not. We refer to this as "fatality" with 1 if the person died and 0 if the person did not die. For this study, we consider race as the protected attribute. While the dataset includes 5 race values, we restrict our analysis to the binary case of White vs Black. We begin by computing the contingency table for the dataset. We set the total number of records in synthetic datasets to 40K. For the various scenarios, we vary the proportion of Whites from 10% to 90% and the outcome prevalence from 10% to 90% as well, both with steps of 10%. This creates a grid of possible scenarios. IPF is then used to determine the cell counts based on these proportions. For simplicity, we use bootstrapping with replacement as the conditional generator such that the new dataset proportions match those identified by IPF.

For the second case study, we study the Adult Income dataset [Kohavi and Becker 1996] which includes details about individuals and whether they earn more than $50K or not. This dataset is also one of the most commonly used datasets for fairness analysis. The dataset includes several features, and we use gender as the protected attribute for this analysis. Similar to the first case study, we traverse the proportions of the protected attribute and outcome to generate a grid of synthetic scenarios using IPF and bootstrapping. A recent study revealed the limitations of this dataset [Ding et al. 2021] and introduced datasets for fairness using additional data sources. In contrast, rather than relying on the availability of additional sources of data, we use synthetic data generation with distribution shifts to create new datasets for fairness evaluation from the available dataset itself.

Reweighing [Calders et al. 2009] improves fairness by weighting the samples of the given dataset such that the data is more fair before the ML model is trained. We apply reweighing on the various synthetic datasets and then train Random Forest (RF) models on them. For the fairness metric, we evaluate Equalized Odds (EO) [Hardt et al. 2016], which measures the maximum of the differences in the True Positive Rate (TPR) and False Positive Rate (FPR) between the subgroups of a protected attribute. We define the fairness region to be the interval $[0, 0.1]$ with 0 indicating exact fairness, as done in a previous study as in [Bhanot et al. 2022]. If the two subgroups are defined by $sub1$ and $sub2$, then EO is defined by:

$$EO = \max(|FPR_{sub1} - FPR_{sub2}|, |TPR_{sub1} - TPR_{sub2}|). \quad (1)$$

The observed results, i.e. Reweighing, are compared with models generated without applying reweighing, i.e. Baseline, to highlight

scenarios where reweighing worked and where it did not. For each scenario, the data is split into 70-30 train-test split. A Random Forest Classifier is trained on the training data and then evaluated for fairness and balanced accuracy on the test data. This experiment is repeated 10 times with changing seed values and the 95% confidence interval along with the averaged score is recorded. The Equalized Odds for the two datasets are shown in Figures 2 and 3 while additional fairness metric results and balanced accuracy scores are available in appendix.

3 PRELIMINARY RESULTS

3.1 Case study 1: MIMIC-III dataset

Our first case study takes a subset of the MIMIC-III dataset with race as the protected attribute. We use IPF and bootstrapping with replacement to create a grid of scenarios with compound shifts. Figure 2 (a) shows the EO results of reweighing applied to all scenarios. We observe that reweighing consistently produces fair models across the majority of scenarios but still struggles amongst a few. For example, when the non-fatality to fatality rate is 70-to-30, reweighing isn't able to reduce bias below the threshold of 0.10, when Whites are in the majority.

In Figure 2 (b), we compare the EO results for reweighing to the baseline of no fairness mitigation, when the Black-to-White proportion is set as 20-to-80. Generally, reweighing improves fairness compared to baseline, as is clear by the orange line lying closer to 0 than the blue line. However, when the non-fatality rate is 50% or 70%, reweighing does not reduce bias below a level of 0.1 as seen by the orange line extending outside the green region of fairness (0 to 0.1). This shows that even while reweighing generally works for the protected attribute of race in this dataset, it doesn't always resolve unfairness under distribution shift.

3.2 Case study 2: Adult Income dataset

Figure 3 (a) shows EO results for Adult Income with gender (Male vs Female) as the protected attribute. Reweighing struggles to improve the fairness across almost all the synthetic scenarios generated. Clearly, reweighing is not an ideal fairness mitigation technique for many scenarios.

Figure 3 (b) compares reweighing to the baseline for Female-to-Male ratio set to 50-to-50. Reweighing has better fairness when 10% to 70% of people earn income less than or equal to $50K, but performs worse than baseline in other cases. Even in the cases where fairness is improved, the scores still lie outside the region of fairness. While the average EO score of reweighing (orange) for 60% to 70% lies within the region of fairness, the 95% confidence interval shows that this may not be the case when experiments are repeated. Thus, reweighing may not be a preferred fairness mitigation technique for the gender attribute in this dataset when Equalized Odds is of interest.

4 DISCUSSION AND CONCLUSION

With growing needs to address ML fairness, fairness mitigation techniques provide a viable solution. However, current evaluation of such techniques is not robust and thus hinders model translation in real-world scenarios. In healthcare, fairness considerations are especially important as decisions can be life-altering.

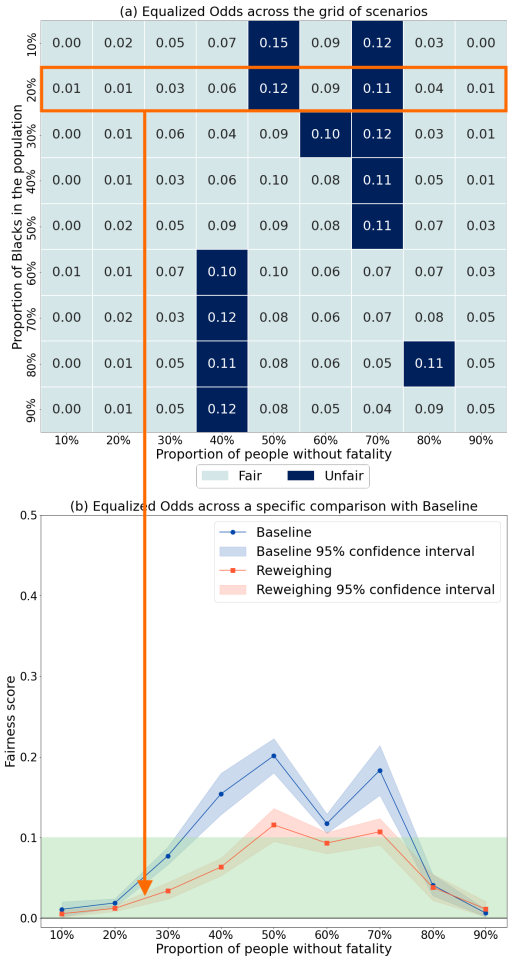

Figure 2: Equalized Odds (EO) scores observed for the race protected attribute in MIMIC-III. (a) EO scores observed after applying reweighing in various synthetic scenarios, where light blue implies relative fairness while dark blue implies unfairness. (b) Comparing EO scores of reweighing (orange) with baseline (blue) ML model when Black-to-White distribution is set to 20-to-80. Bands indicate the 95% confidence interval.

We proposed the design of an auditor that uses synthetic data generation as means for realizing a grid of distribution shifts from a given dataset. The resulting suite of shifted datasets is then used for stress-testing fairness mitigation techniques to identify and empirically quantify their applicability. In two case studies based on MIMIC-III and Adult Income, we used the auditor to generate new dataset scenarios with shifts, and then assessed the reweighing mitigation technique on all scenarios. We found that reweighing is not a sufficient mitigation technique to address biases in these shifted datasets. While these results are preliminary, the proposed design holds several benefits outlined below.

Addressing the problem of limited datasets: Lack of good-quality datasets limits testing of ML models and fairness mitigation

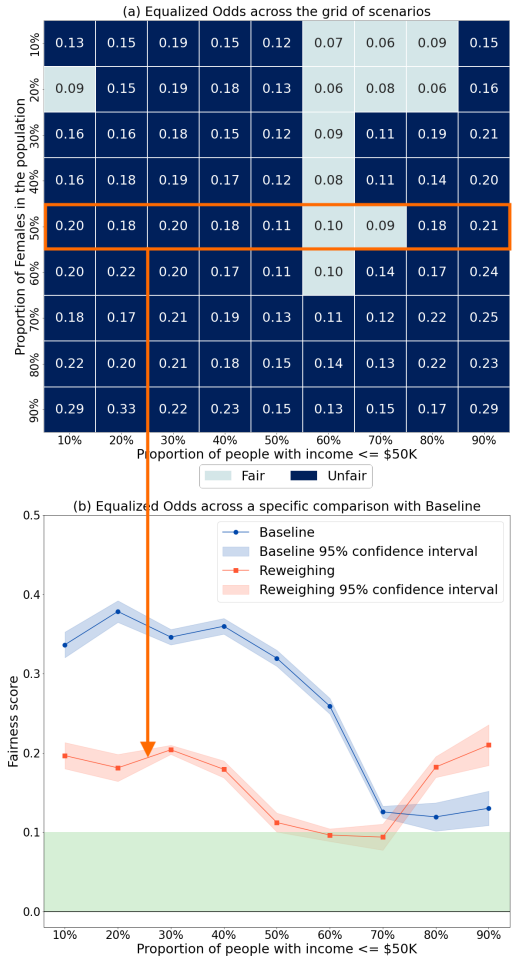

Figure 3: Equalized Odds (EO) scores observed for the gender protected attribute in Adult Income. (a) EO scores observed after applying reweighing in various synthetic scenarios, where light blue implies relative fairness while dark blue implies unfairness. (b) Comparing EO scores of reweighing (orange) with baseline (blue) ML model when Female-to-Male distribution is set to 50-to-50. Bands indicate the 95% confidence interval.

techniques. The proposed auditor design addresses this by using synthetic data generation to generate several datasets from an original dataset by introducing distribution shifts. A single dataset becomes many to evaluate a given technique. Many choices of synthetic data generators are available to the evaluator, such as bootstrapping, copulas, GANs, etc., depending on the desired features of the synthetic data. Exploring different data generators is an interesting direction for future research.

Releasing data while preserving privacy: Using privacy-preserving synthetic data generators like DPGAN [Xie et al. 2018], HealthGAN [Yale et al. 2020a,b] etc., a private dataset can be used as an input to the auditor to create synthetic datasets with distribution shifts. These datasets can then be released along with published research to aid reproducibility and ensure data availability. This will be useful in healthcare where real data release is often restricted under regulations and contracts.

Generating comprehensive reports: The auditor can report scores across a wide variety of metrics. Our prototype interactive app allows the evaluator to assess performance of the mitigation technique visually and quantitatively for many metrics. This is essential as the choice of fairness and utility metric often depends on the application and domain.

4.1 Facilitating ethical deployment

While developing models and evaluating them has been simplified by using several open-source libraries, there is a need to strengthen the real-world value of such evaluation. Evaluations relying on limited, overly-used datasets for fairness provide little information on the applicability of models and techniques in real-world scenarios. For example, directly applying a model designed for a region with a majority White population to another region with a majority Black population may fall short ethically as the model is not trained to handle such a shift. Very different performance and undesired biases can thus result. Deploying an auditor like the one described here can enable a more comprehensive look at the model's performance and provide insights into where it should be deployed.

In this paper, we proposed the idea of using an auditor to stress-test fairness mitigation techniques as a means of robust evaluation. In the future, we plan to perform a comprehensive analysis of the design using different datasets, synthetic data generators and fairness mitigation techniques to strengthen the impact of this idea.

REFERENCES

A Agarwal, A Beygelzimer, Mv Dudík, J Langford, and H Wallach. 2018. A Reductions Approach to Fair Classification. In Intl Conf on Machine Learning. PMLR, 60–69.

R K E Bellamy, K Dey, M Hind, S C Hoffman, S Houde, K Kannan, P Lohia, J Martino, S Mehta, A Mojsilovic, S Nagar, K Natesan Ramamurthy, J Richards, D Saha, P Sattigeri, M Singh, K . Varshney, and Y Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. arXiv:1810.01943

K Bhanot, I Baldini, D Wei, J Zeng, and K P Bennett. 2022. Downstream Fairness Caveats with Synthetic Healthcare Data. arXiv:2203.04462

J Buolamwini and T Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Conf on Fairness, Accountability and Transparency. PMLR, 77–91.

T Calders, F Kamiran, and M Pechenizkiy. 2009. Building Classifiers with Independency Constraints. In 2009 IEEE Intl Conf on Data Mining Workshops. IEEE, 13–18.

Cntrs for Disease Cont and Prev. 2018. Health Insurance Portability and Accountability Act of 1996 (HIPAA). https://www.cdc.gov/phlp/publications/topic/hipaa.html

A Coston, K Natesan Ramamurthy, D Wei, K R Varshney, S Speakman, Z Mustahsan, and S Chakraborty. 2019. Fair Transfer Learning with Missing Protected Attributes. ACM, New York, NY, USA.

F Ding, M Hardt, J Miller, and L Schmidt. 2021. Retiring Adult: New Datasets for Fair Machine Learning. In Adv in Neural Info Proc Sys, M Ranzato, A Beygelzimer, Y. Dauphin, P S Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 6478–6490.

European Parliament and of the Council (2016, Apr. 27). 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). Official Journal, L119 (May 2016), 1–88.

M Hardt, E Price, and N Srebro. 2016. Equality of Opportunity in Supervised Learning. arXiv:610.02413

Alistair E W Johnson, T J Pollard, L Shen, L H Lehman, M Feng, Benj Ghassemi, Mand Moody, P Szolovits, L Anthony Celi, and R G Mark. 2016. MIMIC-III, a Freely Accessible Critical Care Database. Scientific Data 3, 1 (2016), 1–9.

F Kamiran, A Karim, and X Zhang. 2012. Decision Theory for Discrimination-aware Classification. In IEEE 12th Intl Conf on Data Mining. IEEE, 924–929.
R Kohavi and B Becker. 1996. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/adult
N Lomax and P Norman. 2016. Estimating Population Attribute Values in a Table: "Get Me Started in" Iterative Proportional Fitting. The Professional Geographer 68, 3 (2016), 451–461.
K Makhlouf, S Zhioua, and C Palamidessi. 2021. On the Applicability of Machine Learning Fairness Notions. 23, 1 (2021).
P Norman. 1999. Putting Iterative Proportional Fitting on the Researcher's Desk. (1999).
J Schrouff, N Harris, O Koyejo, I Alabdulmohsin, E Schnider, K Opsahl-Ong, A Brown, S Roy, D Mincu, C Chen, A Dieng, Y Liu, V Natarajan, A Karthikesalingam, K Heller, S Chiappa, and A D'Amour. 2022. Maintaining Fairness Across Distribution Shift: Do We Have Viable Solutions for Real-world Applications? arXiv:2202.01034
S Shankar, Y Halpern, E Breck, J Atwood, J Wilson, and D Sculley. 2017. No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. arXiv:1711.08536
L Xie, K Lin, S Wang, F Wang, and J Zhou. 2018. Differentially Private Generative Adversarial Network. CoRR abs/1802.06739 (2018).
A Yale, S Dash, K Bhanot, I Guyon, J S. Erickson, and K P Bennett. 2020a. Synthesizing Quality Open Data Assets from Private Health Research Studies. In Business Information Systems Workshops, Witold Abramowicz and Gary Klein (Eds.). Springer Intl Pub, Cham, 324–335.
A Yale, S Dash, R Dutta, I Guyon, A Pavao, and K P Bennett. 2020b. Generation and Evaluation of Privacy Preserving Synthetic Health Data. Neurocomputing 416 (2020), 244–255.

A UTILITY SCORES

While fairness mitigation techniques aim to resolve biases in the data and model, they should not do so at the expense of losing model performance. Thus, the "utility" of the trained model should be high for it to be useful for real-world applications. In our experiments, we measure utility using test balanced accuracy scores. Additionally, the auditor reporting can be extended to include other utility metrics as well.

We evaluated the test balanced accuracy scores corresponding to the Equalized Odds scores calculated for race attribute in MIMIC-III and gender attribute in Adult Income dataset. As before, each experiment is repeated 10 times and the averaged value along with the 95% confidence interval is reported.

A.1 Case study 1: MIMIC-III dataset

Figure 4 shows the test balanced accuracy scores when changing the proportion of population without fatality from 10% to 90% with Black-to-White population set as 20-to-80. We observe that after applying reweighing, the utility doesn't change much. The confidence intervals for all the scenarios overlap indicating negligible loss in utility. However, in extreme cases (10% and 90%), the model had almost 0.5 balanced accuracy indicating that the model struggled to learn much from the data when the datasets were extreme, with and without reweighing.

A.2 Case study 2: Adult Income dataset

Figure 5 shows the test balanced accuracy scores when changing the proportion of population who earn less than or equal to $50K with Female-to-Male population set as 50-to-50. In contrast to MIMIC-III models, we observe that the Random Forest Classifier had better test balanced accuracy scores throughout, as all values are above 0.5. Furthermore, in all the observed cases, the utility of the model after reweighing was lower than the baseline models. This is concerning, as even this reduction in utility isn't accompanied by a complete removal of bias from the data as measured by Equalized Odds.

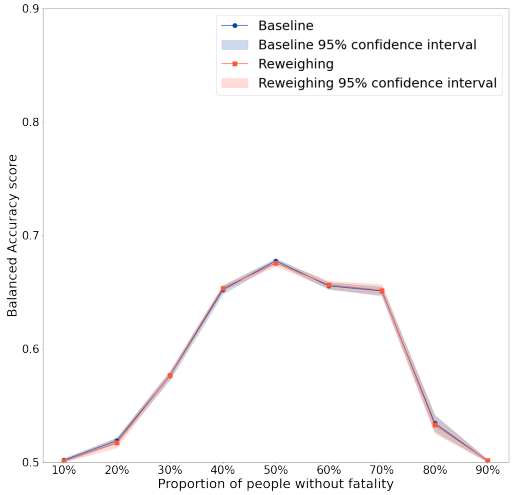

Figure 4: Test Balanced Accuracy scores observed for training Random Forest Classifier on MIMIC-III. The blue line with circles describes the scores for baseline model with 95% confidence interval while the orange line with squares describes the scores for model generated after reweighing with 95% confidence interval.

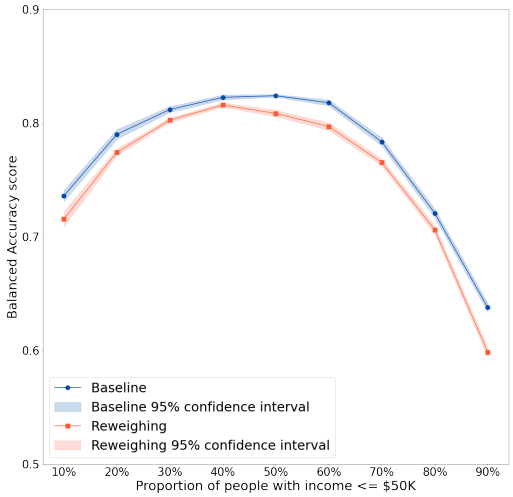

Figure 5: Test Balanced Accuracy scores observed for training Random Forest Classifier on Adult Income. The blue line with circles describes the scores for baseline model with 95% confidence interval while the orange line with squares describes the scores for model generated after reweighing with 95% confidence interval.

B FAIRNESS SCORES

To perform a robust evaluation, we also measured fairness using Disparate Impact (DI). DI is defined as the ratio of selection rates as shown in Equation 2. Put simply, DI is ratio of the probability Pr of predicting the positive label $\hat{Y} = 1$ for a given protected group

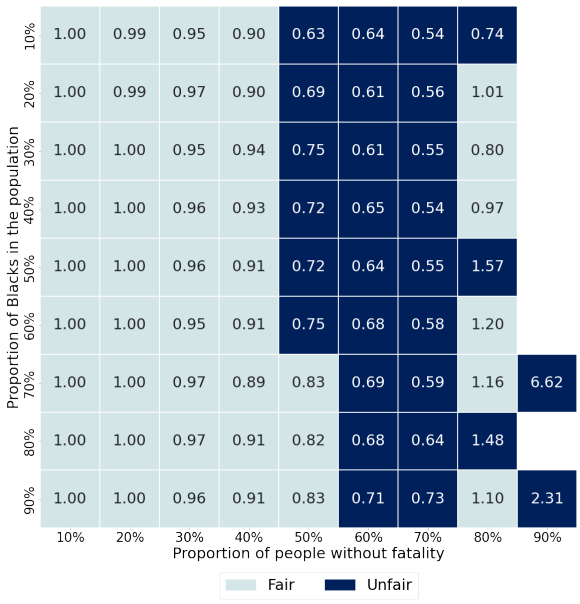

Figure 6: Disparate Impact (DI) observed for the race protected attribute in MIMIC-III after applying reweighing in various synthetic scenarios where light blue implies fairness while dark blue implies unfairness scores.

$g(x) = 1$ to the probability Pr of predicting the positive label $\hat{Y} = 1$ for the rest of the population $g(x) = 0$.

$$DI = \frac{Pr(\hat{Y} = 1|g(x) = 1)}{Pr(\hat{Y} = 1|g(x) = 0)} \tag{2}$$

B.1 Case study 1: MIMIC-III dataset

From Figure 6, we observe that reweighing performs well in some scenarios (light blue) but does poorly for others (dark blue). In the most extreme cases, DI score is either 1 or Nan, indicating the model struggled to learn the data. This was observed in utility scores before as seen in Figure 4, where the balanced accuracy value was close to 0.5. For the other scenarios, reweighing did not perform as well. Thus, if DI is an important metric for the application of this specific dataset with race protected attribute, then reweighing might not be the ideal fairness mitigation technique.

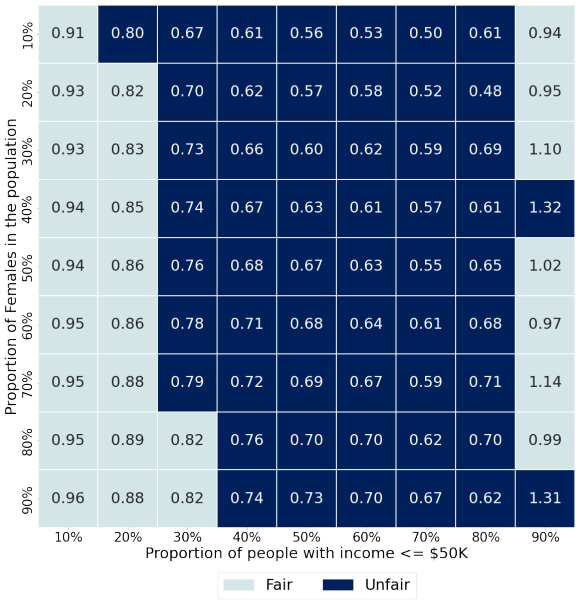

Figure 7: Disparate Impact (DI) observed for the gender protected attribute in Adult Income after applying reweighing in various synthetic scenarios where light blue implies fairness while dark blue implies unfairness scores.

B.2 Case study 2: Adult Income dataset

Figure 7 shows the DI scores for the various synthetic scenarios generated from Adult Income dataset. We observe that for majority cases, the DI scores fall outside the range of fairness as indicated by dark blue. In the extreme cases, the fairness as measured by DI lies within the fair region, as indicated by light blue. However, overall, we can conclude that reweighing struggled to remove the bias from the dataset completely and thus, might not be an ideal technique for removing bias for both Equalized Odds and Disparate Impact fairness.